

# Spatial Neutral to the Right Species Sampling Mixture Models

LANCELOT F. JAMES<sup>1</sup>

*The Hong Kong University of Science and Technology*

This paper describes briefly how one may utilize a class of species sampling mixture models derived from Doksum's (1974) neutral to the right processes. For practical implementation we describe an ordered/ranked variant of the generalized weighted Chinese restaurant process.

## 1 Introduction

The field of Bayesian nonparametric statistics essentially involves the idea of assigning prior and posterior distributions over spaces of probability measures or more general measures. That is, similar to the classical parametric Bayesian idea of assigning priors to an unknown parameter, say  $\theta$ , which lies in a Euclidean space, one views, for instance, an unknown cumulative distribution function, say  $F(t)$ , as being a stochastic process. More generally for an unknown probability measure  $P$ , a Bayesian views it as a random probability measure. This is currently a well-developed and active area of research that has links to a variety of areas where Lévy and more general random processes are commonly used. However, as discussed in Doksum and James (2004), in the late 1960's, noting the high activity and advance in nonparametric statistics, David Blackwell and others wondered how one could assign priors which were both flexible and tractable. Arising from these questions were two viable answers which till this day remain at the cornerstone of Bayesian nonparametric statistics.

Ferguson (1973, 1974) proposed the use of a Dirichlet process prior[see also Freedman (1963)]. For this prior if  $P$  is a probability on some space  $\mathcal{X}$ , and  $(B_1, \dots, B_k)$  is a measurable partition of  $\mathcal{X}$ , then  $P(B_1), \dots, P(B_k)$  has a Dirichlet distribution. Moreover, the posterior distribution of  $P$  given a sample  $\mathbf{X} = (X_1, \dots, X_n)$  is also a Dirichlet process. For a specified probability measure  $H$  and a scalar  $\theta > 0$ , one can say that  $P \stackrel{d}{=} P_{\theta H}$  is a Dirichlet process with shape parameter  $\theta H$ , if the Dirichlet distributions discussed above have parameters given by  $\mathbb{E}[P(A_i)] = \theta H(A_i)$ . Following this, Doksum (1974) introduced the class of Neutral to the Right (NTR) random probability measures on the real line. For these models if  $P$  is a distribution on the real line, then for each partition  $B_1, \dots, B_k$ , with  $B_j = (s_{j-1}, s_j]$ ,  $j = 1, \dots, k$ ,  $s_0 = -\infty, s_k = \infty$ ,  $s_i < s_j$  for  $i < j$ ;  $P(B_1), \dots, P(B_k)$  is such that  $P(B_i)$  has the same distribution as  $V_i \prod_{j=1}^{i-1} (1 - V_j)$ , where  $V_1, \dots, V_2, \dots$  is a collection of independent non-negative random variables. This represents a remarkably rich choice of models defined by specifying different distributions for the  $V_i$ . Notably if  $V_i$  is chosen to be beta random variable with parameters  $(\alpha_i, \beta_i)$  and  $\beta_i = \sum_{j=1}^{k-1} \alpha_j$ , then this gives the Dirichlet process as described in Doksum (1974). Doksum (1974) shows that if  $P$  is a NTR distribution then the posterior distribution of  $P$  give a sample  $X_1, \dots, X_n$  is also an NTR. Subsequently, Ferguson and Phadia (1979), showed that this type of conjugacy property extends to the case of right censored survival models. This last fact coupled with the subsequent related works of Hjort (1990), Kim (1999), Lo (1993) and Walker and Muliere (1997) have popularized the usage of NTR processes in models related to survival and event history analysis.

<sup>1</sup>AMS 2000 subject classifications. Primary 62G05; secondary 62F15.

Corresponding authors address. The Hong Kong University of Science and Technology, Department of Information Systems and Management, Clear Water Bay, Kowloon, Hong Kong. lance@ust.hk

Keywords and phrases. Chinese Restaurant process, Dirichlet process, Lévy processes, Neutral to the right processes, Species sampling models.

Despite these attractive points, the usage of NTR processes in more complex statistical models, such as mixture models, has been notably absent. This is in contrast to the Dirichlet process which, coupled with the advances in MCMC and other computational procedures, is regularly used in nonparametric or semi-parametric statistical models. The theoretical framework for Dirichlet process mixture models can be traced back to Lo (1984) who proposed to model a density as a convolution mixture model of a known kernel density  $K(y|x)$  and a Dirichlet process  $P$  as,

$$(1) \quad f(y|P) = \int_{\mathcal{X}} K(y|x)P(dx).$$

This may be equivalently expressed in terms of a missing data model where for a sample  $\mathbf{Y} = (Y_1, \dots, Y_n)$  based on (1), one has  $Y_1, \dots, Y_n | \mathbf{X}, P$  are such that  $Y_i$  are independent with distributions  $K(\cdot|X_i)$ ,  $X_i|P$  are iid  $P$  and  $P$  is a Dirichlet process. It is clear that the description of the posterior distribution of  $P$  and related quantities is much more complex than in the setting discussed in Ferguson (1973). However, Lo (1984) shows that its description is facilitated by the descriptions of the posterior distribution of  $P|\mathbf{X}$ , given by Ferguson (1973) and the exchangeable marginal distribution of  $\mathbf{X}$  discussed in Blackwell and MacQueen (1973). Blackwell and MacQueen describe the distribution via what is known as the Blackwell-MacQueen Pólya urn scheme where  $\mathbb{P}(X_1 \in A) = H(A)$  and for  $n > 1$

$$(2) \quad \mathbb{P}(X_n \in \cdot | X_1, \dots, X_{n-1}) = \frac{\theta}{\theta + n - 1} H(\cdot) + \frac{1}{\theta + n - 1} \sum_{j=1}^{n-1} \delta_{X_j}(\cdot).$$

Note that (2) clearly indicates that there are ties among  $(X_1, \dots, X_n)$  and that the  $n(\mathbf{p}) \leq n$  unique values, say  $X_1^*, \dots, X_{n(\mathbf{p})}^*$  are iid with common distribution  $H$ . Letting  $\mathbf{p} = \{C_1, \dots, C_{n(\mathbf{p})}\}$  denote a partition of the integers  $\{1, \dots, n\}$ , where one can write  $C_j = \{i : X_i = X_j^*\}$ , with size  $n_j = |C_j|$  for  $j = 1, \dots, n(\mathbf{p})$ . This leads to the following important description of the distribution of  $\mathbf{X}$ ,

$$\pi(d\mathbf{X}|\theta H) = \text{PD}(\mathbf{p}|\theta) \prod_{j=1}^{n(\mathbf{p})} H(dX_j^*)$$

where

$$\text{PD}(\mathbf{p}|\theta) = \frac{\theta^{n(\mathbf{p})} \Gamma(\theta)}{\Gamma(\theta + n)} \prod_{j=1}^{n(\mathbf{p})} (n_j - 1)! := p_{\theta}(n_1, \dots, n_{n(\mathbf{p})})$$

is a variant of Ewens sampling formula [see Ewens (1972) and Antoniak (1974)], often called the Chinese restaurant process. It can be interpreted as  $\mathbb{P}(C_1, \dots, C_{n(\mathbf{p})}) = p_{\theta}(n_1, \dots, n_{n(\mathbf{p})})$  where  $p_{\theta}$ , being symmetric in its arguments, is the most notable example of an *exchangeable partition probability function* (EPPF) [see Pitman (1996)]. It is easily seen that a Dirichlet Process with shape  $\theta H$  is characterized by the pair  $(p_{\theta}, H)$ . Letting  $p(n_1, \dots, n_k)$ , for  $n(\mathbf{p}) = k$ , denote an arbitrary EPPF, Pitman (1996) shows that the class of random probability measures whose distribution is completely determined by the pair  $(p, H)$  must correspond to the class of *species sampling random probability measures*. General species sampling random probability measures constitute all random probability measures that can be represented as

$$(3) \quad P(\cdot) = \sum_{i=1}^{\infty} P_i \delta_{Z_i}(\cdot) + (1 - \sum_{k=1}^{\infty} P_k) H(\cdot)$$

where  $0 \leq P_i < 1$  are random weights such that  $0 < \sum_{i=1}^{\infty} P_i \leq 1$ , independent of the  $Z_i$  which are iid with some non-atomic distribution  $H$ . Furthermore the law of the  $(P_i)$  is determined by the EPPF  $p$ . Noting these points Ishwaran and James (2003) described the class of species sampling mixture models by replacing a Dirichlet process in (1) with  $P$  specified by (3). See also Müller and Quintana (2004).

Except for the special case of the Dirichlet process, NTR processes are not species sampling models and this is one of the factors which makes analysis a bit more difficult. Nonetheless, James (2003, 2006) was able to extend the definition of NTR processes to a class of random probability measures on more general spaces, which he called Spatial NTR processes. Additionally a tractable description of the marginal distribution of this class of models was obtained. These two ingredients then allow for the implementation of NTR mixture models. Our goal in this note is not to describe the mechanisms for a full-blown NTR mixture model, as this requires much more overhead, but rather mixture models based on species sampling models which are derived from NTR processes. James (2003, 2006) introduced and calls these *NTR species sampling models*. Quite specifically, though the NTR processes are not species sampling models they produce EPPF's  $p$  that, along with the specification of  $H$ , are uniquely associated with an NTR species sampling model. This produces a very rich and flexible class of random priors that are a bit simpler analytically than NTR processes. An interesting fact is that this class contains the two-parameter  $(\alpha, \theta)$  Poisson-Dirichlet random probability measures for parameters  $0 \leq \alpha < 1$  and  $\theta > 0$ . That is the Dirichlet process and a class of random probabilities defined by normalizing a stable law process and further power tempering the stable law distribution, which are discussed in Pitman (1996) and Pitman and Yor (1997). Implementations of these latter models, being quite special, may be treated by computational procedures involving random partitions discussed in Ishwaran and James (2003) or by the methods in Ishwaran and James (2001). Here we will discuss a ranked weighted Chinese restaurant procedure which applies more generally.

**2. NTR and related processes.** Let  $F(t)$  denote an NTR cumulative distribution function on the positive real line. Additionally, let  $S(t) = 1 - F(t)$  denote a survival function. Doksum (1974) Theorem 3.1 shows that  $F$  is an NTR process if and only if it can be represented as

$$(4) \quad F(t) = 1 - e^{-Y(t)}$$

where  $Y(t)$  is an independent increment process which is non-decreasing and right continuous almost surely and furthermore  $\lim_{t \rightarrow \infty} Y(t) = \infty$  and  $\lim_{t \rightarrow -\infty} Y(t) = 0$  almost surely. In other words  $Y$  belongs to the class of positive Lévy processes. We shall suppose hereafter that  $T$  is a positive random variable such that given  $F$  its distribution function is  $F$  where  $F$  is an NTR process. Then  $T$  has an interpretation as a survival time with “conditional” survival distribution  $S(t) = 1 - F(t) := P(T > t|F)$ . It is evident from (4) that the distribution of  $F$  is completely determined by the law of  $Y$  which is determined by its Laplace transform

$$E \left[ e^{-\omega Y(t)} \right] = e^{-\int_0^t \phi(\omega|s) \Lambda_0(ds)} := E[(S(t))^\omega]$$

where  $\phi(\omega|s)$  is equal to

$$(5) \quad \int_0^\infty (1 - e^{-v\omega}) \tau(dv|s) = \int_0^1 (1 - (1-u)^\omega) \rho(du|s) = \int_0^1 \omega(1-u)^{\omega-1} \left[ \int_u^1 \rho(dv|s) \right] du$$

$\tau$  and  $\rho$  are Lévy densities on  $[0, \infty]$  and  $[0, 1]$  respectively which are in correspondence via the mapping  $y \rightarrow 1 - e^{-y}$ . Without loss of generality we shall assume that  $\int_0^1 u \rho(du|s) = 1$  for each fixed  $s$ , which implies that  $\phi(\omega|s) = 1$ . Hence we have that

$$E[S(t)] = e^{-\Lambda_0(t)} = 1 - F_0(t)$$

where  $F_0$  represents one's prior belief about the true distribution and  $\Lambda_0(dt) = F_0(dt)/S_0(t-)$  is its corresponding cumulative hazard with  $S_0(t-) = 1 - F_0(t-) = \mathbb{P}(T \geq t)$ .

Note that for each fixed  $s$ ,  $\phi(\omega|s)$  corresponds to the log Laplace transform of an infinitely-divisible random variable. It follows that different specifications for  $\tau$  or equivalently  $\rho$  lead to different NTR processes. When  $\tau$  and  $\rho$  do not depend on  $s$ ,  $F$ ,  $Y$  and all relevant functionals are said to be *homogeneous*. We also apply this name to  $\tau$  and  $\rho$ . Additionally  $\phi(\omega|s)$  specializes to

$$\phi(\omega) := \int_0^\infty (1 - e^{-v\omega}) \tau(dv) = \int_0^1 (1 - (1-u)^\omega) \rho(du) = \int_0^1 \omega(1-u)^{\omega-1} \left[ \int_u^1 \rho(dv) \right] du.$$

Consider now the cumulative hazard process of  $F$ , say  $\Lambda$ , defined by  $\Lambda(dt) = F(dt)/S(t-)$ . The idea of Hjort (1990) was to work directly with  $\Lambda$  rather than  $F$ . He showed importantly that if one specified  $\Lambda$  to be a positive completely random measure on  $[0, 1]$ , whose law is specified by the Laplace transform

$$\mathbb{E}[e^{-\omega\Lambda(t)}] = e^{-\int_0^t \psi(\omega|s)\Lambda_0(ds)}$$

where  $\psi(\omega|s) := \int_0^1 (1 - e^{-u\omega})\rho(du|s)$ , then  $F$  and  $S$  must be NTR processes specified by (5). James (2003, 2006) shows that one can extend the definition of an NTR process to a spatial NTR process on  $[0, \infty] \times \mathcal{X}$  by working with the concept of a random hazard measure, say  $\Lambda_H(dt, dx)$ .  $\Lambda_H$  is a natural extension of  $\Lambda$  in the sense that  $\Lambda_H(dt, \mathcal{X}) = \Lambda(dt)$  and is otherwise specified by replacing the intensity  $\rho(du|s)\Lambda_0(ds)$  by  $\rho(du|s)\Lambda_0(ds, dx)$ , where,

$$\Lambda_0(ds, dx) = H(dx|s)\Lambda_0(ds)$$

is a hazard measure and  $H(\cdot|s)$  may be interpreted as the conditional distribution of  $X|T = s$ . A Spatial NTR process (SPNTR) is then defined as

$$(6) \quad P_S(dt, dx) = S(t-)\Lambda_H(dt, dx)$$

The SPNTR in (6) has marginals such that  $P_S(dt, d\mathcal{X}) = F(dt)$  is an NTR and

$$(7) \quad P_S([0, \infty), dx) = \int_0^\infty S(t-)\Lambda_H(ds, dx),$$

represents an entirely new class of random probability measures.

## 2.1 NTR species sampling models

NTR species sampling models arise as a special case of (7) by setting  $H(dx|s) := H(dx)$ . Here we will further work only with the class of homogeneous processes and hence we will additionally choose  $\rho(du|s) = \rho(du)$ . Thus an NTR species sampling model is of the form

$$P_{\rho, H}(dx) = \int_0^\infty S(s-)\Lambda_H(ds, dx) = \sum_{k=1}^\infty P_k \delta_{Z_k}(dx).$$

Furthermore, if  $P \stackrel{d}{=} P_{\rho, H}$  then we denote its law as  $\mathcal{P}(\cdot|\rho, H)$ . It follows that for practical usage in mixture models one needs a tractable description of the corresponding EPPF, say  $p_\rho$ . However, before we do that we will need to introduce additional notation which connects  $p_\rho$  with the NTR process. If we suppose that  $X_1, \dots, X_n|P_{\rho, H}$  are iid with distribution  $P_{\rho, H}$ , then these points come from a description of the  $n$  conditionally independent pairs  $(T_1, X_1), \dots, (T_n, X_n)|P_S$  where  $(T_i, X_i)$  are iid  $P_S$ , such that  $T_i$  are iid  $F$ , where  $F$  is an NTR, and  $X_i$  are iid  $P_{\rho, H}$ . Here  $P_S$  must be specified by the intensity  $\rho(du)\Lambda_0(ds)H(dx)$ . Now if one denotes the  $n(\mathbf{p})$  unique pairs as  $(T_j^*, X_j^*)$  for  $j=1, \dots, n(\mathbf{p})$ , then one may simply set each  $C_j = \{i : T_i = T_j^*\}$ . Furthermore we define  $T_{(1:n)} > T_{(2:n)} > \dots > T_{(n(\mathbf{p}):n)} > 0$  to be the ordered values of the unique values  $(T_j^*)_{j \leq n(\mathbf{p})}$ . Hence we can define  $\mathbf{p}$  by setting  $C_j := \{i : T_i = T_j^*\}$ , and define  $\mathbf{m} = \{D_1, \dots, D_{n(\mathbf{p})}\}$  with cells  $D_j = \{i : T_i = T_{(j:n)}\}$  with cardinality  $d_j = |D_j|$ . It is evident that given a partition  $\mathbf{p} = \{C_1, \dots, C_{n(\mathbf{p})}\}$ ,  $\mathbf{m}$  takes its values over the symmetric group, say  $\mathcal{S}_{n(\mathbf{p})}$ , of all  $n(\mathbf{p})!$  permutations of  $\mathbf{p}$ . Let  $R_{j-1} = \bigcup_{k=1}^{j-1} D_k := \{i : T_i > T_{(j:n)}\}$  with cardinality  $r_{j-1} = \sum_{k=1}^{j-1} d_k$ . Then, in terms of survival analysis, the quantities  $d_j$  and  $r_j = d_j + r_{j-1}$  have the interpretation as the number of deaths at time  $T_{(j:n)}$ , and the number at risk at time  $T_{(j:n)}$ , respectively. See James (2006) for some further elaboration. Now from James (2003, 2006) it follows that

$$(8) \quad \pi_\rho(\mathbf{p}) = p_\rho(n_1, \dots, n(\mathbf{p})) = \sum_{\mathbf{m} \in \mathcal{S}_{n(\mathbf{p})}} \frac{\prod_{j=1}^{n(\mathbf{p})} \kappa_{d_j, r_{j-1}}(\rho)}{\prod_{j=1}^{n(\mathbf{p})} \phi(r_j)}$$

where,

$$\kappa_{d_j, r_{j-1}}(\rho) = \int_0^1 u^{d_j} (1-u)^{r_{j-1}} \rho(du).$$

The form of the EPPF is in general not directly tractable. However by augmentation one sees that the distribution of  $\mathbf{m}$  is given by

$$(9) \quad \pi_\rho(\mathbf{m}) = \frac{\prod_{j=1}^{n(\mathbf{p})} \kappa_{d_j, r_{j-1}}(\rho)}{\prod_{j=1}^{n(\mathbf{p})} \phi(r_j)}$$

and has a nice product form. This suggests that one can work with a joint distribution of  $(\mathbf{X}, \mathbf{m})$  given by

$$\pi_\rho(\mathbf{m}) \prod_{j=1}^{n(\mathbf{p})} H(dX_j^*).$$

Related to this, James (2006) shows that a prediction rule of  $X_{n+1}|\mathbf{X}, \mathbf{m}$  is given by

$$\mathbb{P}(X_{n+1} \in dx | \mathbf{X}, \mathbf{m}) = (1 - \sum_{j=1}^{n(\mathbf{p})} p_{j:n}) P_0(dx) + \sum_{j=1}^{n(\mathbf{p})} p_{j:n} \delta_{X_j^*}(dx),$$

with  $(1 - \sum_{j=1}^{n(\mathbf{p})} p_{j:n}) = \sum_{j=1}^{n(\mathbf{p})+1} q_{j:n}$ , and where

$$p_{j:n} = \frac{\kappa_{d_j+1, r_{j-1}}(\rho) \prod_{l=j+1}^{n(\mathbf{p})} \kappa_{d_l, r_{l-1}+1}(\rho)}{\kappa_{d_j, r_{j-1}}(\rho) \prod_{l=j+1}^{n(\mathbf{p})} \kappa_{d_l, r_{l-1}}(\rho)} \prod_{l=j}^{n(\mathbf{p})} \frac{\phi(r_l)}{\phi(r_l+1)}.$$

and

$$q_{j:n} = \frac{\kappa_{1, r_{j-1}}(\rho)}{\phi(r_{j-1}+1)} \frac{\prod_{l=j}^{n(\mathbf{p})} \kappa_{d_l, r_{l-1}+1}(\rho)}{\prod_{l=j}^{n(\mathbf{p})} \kappa_{d_l, r_{l-1}}(\rho)} \prod_{l=j}^{n(\mathbf{p})} \frac{\phi(r_l)}{\phi(r_l+1)},$$

with  $q_{n(\mathbf{p})+1:n} = \kappa_{1,n}(\rho)/\phi(n+1)$ , are transition probabilities derived from  $\pi_\rho(\mathbf{m})$ . Note that in the calculation of  $\kappa_{1, r_{j-1}}(\rho)$ ,  $r_{j-1}+1$  is to be used rather than  $r_j = r_{j-1} + m_j$ . As an example, consider the choice of a homogeneous beta process [Hjort (1990), see also Ferguson (1974), Ferguson and Phadia (1979) and Gneden (2004)] defined by

$$\rho(du) = \theta u^{-1} (1-u)^{\theta-1}$$

then it is easily seen that  $\phi(r_j) = \sum_{l=1}^{r_j} \theta/(\theta+l-1)$ , and it follows that in this case

$$p_{j:n} = \frac{d_j}{n+\theta} \prod_{l=j}^{n(\mathbf{p})} \frac{\phi(r_l)}{\phi(r_l+1)} \text{ and } q_{j:n} = \frac{1}{n+\theta} \frac{1}{\sum_{i=1}^{r_{j-1}+1} 1/(\theta+i-1)} \prod_{l=j}^{n(\mathbf{p})} \frac{\phi(r_l)}{\phi(r_l+1)}.$$

REMARK 1. Gneden and Pitman (2005a) also obtained the expressions (8) and (9) independent of James (2003, 2006), and in a different context. See James (2006) for more details.

REMARK 2. Related to this, Gneden and Pitman (2005a) [see additionally Gneden and Pitman (2005b)] showed that the EPPF in (8) corresponds to that of the two-parameter  $(\alpha, \theta)$  Poisson-Dirichlet process with parameters  $0 \leq \alpha < 1$  and  $\theta > 0$  if  $\rho := \rho_{\alpha, \theta}$  is chosen such that,

$$\int_u^1 \rho_{\alpha, \theta}(dv) = \frac{\Gamma(\theta+2-\alpha)}{\Gamma(1-\alpha)\Gamma(1+\theta)} u^{-\alpha} (1-u)^\theta.$$

From this, James (2006) deduced that  $P_{\rho_{\alpha,\theta},H} = \sum_{k=1}^{\infty} W_k \prod_{i=1}^{k-1} (1 - W_i) \delta_{Z_k}$  where  $(W_k)$  are independent beta  $(1 - \alpha, \theta + k\alpha)$  random variables independent of the  $(Z_k)$  which are iid  $H$ . That is a two-parameter  $(\alpha, \theta)$  Poisson-Dirichlet process, for  $0 \leq \alpha < 1$  and  $\theta > 0$  can be represented as the marginal probability measure of a spatial NTR process, as described above. See Pitman and Yor (1997) and Ishwaran and James (2001) for more on the stick-breaking representation of the two parameter Poisson-Dirichlet process.

**3. NTR species sampling mixture models.** Now setting  $P = P_{\rho,H}$  in (1) yields a special case of the species sampling models described in Ishwaran and James (2003). That is

$$(10) \quad \int_{\mathcal{X}} K(y|x) P_{\rho,H}(dx) = \int_{\mathcal{X}} \int_0^{\infty} K(y|x) S(s-) \Lambda_H(ds, dx)$$

is called an NTR species sampling models. We look at the situation where  $Y_1, \dots, Y_n | P_{\rho,H}$  are iid with density or pmf (10). This translates into the hierarchical model,

$$(11) \quad \begin{aligned} Y_i | X_i, P &\stackrel{ind}{\sim} K(Y_i | X_i) \text{ for } i = 1, \dots, n \\ X_i | P &\stackrel{iid}{\sim} P \\ P &\sim \mathcal{P}(\cdot | \rho, H) \end{aligned}$$

In principle, since we have a description of the EPPF, the theoretical results and computational procedures described in Ishwaran and James (2003) apply. However as we have noted in general  $\pi_{\rho}(\mathbf{p})$  is not as simple to work with as  $\pi_{\rho}(\mathbf{m})$ . So here we develop results that allows us to sample from a posterior distribution of  $\mathbf{m}$  rather than partitions. We summarize these results in the next proposition

**Proposition 3.1** *Suppose that one has the model specified in (11). Then the following results holds*

- (i) *The distribution of  $X_1, \dots, X_n | \mathbf{Y}, \mathbf{m}$  is such that the unique values  $X_j^*$  for  $j = 1, \dots, n(\mathbf{p})$  are conditionally independent with distributions*

$$\pi(dX_j^* | D_j) \propto H(dX_j^*) \prod_{i \in D_j} K(Y_i | X_j^*).$$

- (ii) *The posterior distribution of  $\mathbf{m} | \mathbf{Y}$  is,*

$$\pi_{\rho}(\mathbf{m} | \mathbf{Y}) \propto \pi_{\rho}(\mathbf{m}) \prod_{j=1}^{n(\mathbf{p})} \int_{\mathcal{X}} \prod_{i \in D_j} K(Y_i | x) H(dx).$$

- (iii) *The posterior distribution of  $\mathbf{p} | \mathbf{Y}$  is*

$$\sum_{\mathbf{m} \in \mathcal{S}_{n(\mathbf{p})}} \pi_{\rho}(\mathbf{m} | \mathbf{Y}) = \pi_{\rho}(\mathbf{p}) \prod_{j=1}^{n(\mathbf{p})} \int_{\mathcal{X}} \prod_{i \in C_j} K(Y_i | x) H(dx).$$

□

From this result one can compute a Bayesian predictive density of  $Y_{n+1} | \mathbf{m}, \mathbf{Y}$  as,

$$l(n) = f(Y_{n+1} | \mathbf{m}, \mathbf{Y}) = \left[ \sum_{j=1}^{n(\mathbf{p})+1} q_{j:n} \right] \int_{\mathcal{X}} K(Y_{n+1} | x) H(dx) + \sum_{j=1}^{n(\mathbf{p})} p_{j:n} \int_{\mathcal{X}} K(Y_{n+1} | x) \pi(dx | D_j).$$

A Bayesian density estimate analogous to Lo (1984) is then to sum this expression relative to the distribution of  $\mathbf{m} | \mathbf{Y}$ .

**Corollary 3.1** *Consider the model in Proposition 3.1, then a Bayesian predictive density estimator of  $Y_{n+1}|\mathbf{Y}$  is given by*

$$\mathbb{E}[f(Y_{n+1}|P)|\mathbf{Y}] = \sum_{\mathbf{p}} \sum_{\mathbf{m} \in S_{n(\mathbf{p})}} f(Y_{n+1}|\mathbf{m}, \mathbf{Y}) \pi_{\rho}(\mathbf{m}|\mathbf{Y})$$

□

### 3.0.1 Ordered/Ranked generalized weighted Chinese restaurant processes

The significance of the expression for the predictive density, is that we can use  $l(n)$  in precisely the same manner as the predictive densities given  $\mathbf{p}, \mathbf{Y}$ , used in Ishwaran and James (2003) [see also Lo, Brunner and Chan (1996)] to construct computational procedures for approximating posterior quantities. In fact, all the major computational procedures for Dirichlet process mixture models, see for instance Escobar (1994) and Escobar and West (1995), utilize some type of predictive density. Here, in analogy to the gWCR algorithms in Lo, Brunner and Chan (1996) and Ishwaran and James (2003), we define a weighted version of the *Ordered/Ranked generalized Chinese restaurant process* developed in James (2003, 2006), to approximate a draw from  $\pi_{\rho}(\mathbf{m}|\mathbf{Y})$  as follows. For each  $n \geq 1$ , let  $\{D_{1:n}, \dots, D_{n(\mathbf{p}):n}\}$ , denote a seating configuration of the first  $n$  customers, where  $D_{j:n}$  denotes the set of the  $n$  customers seated at a table with common rank  $j$ .

- (i) Given this configuration, the next customer  $n+1$  is seated at an occupied table  $D_{j:n}$ , denoting that customer  $n+1$  is equivalent to the  $j$ th largest seated customers, with probability,

$$(12) \quad \frac{p_{j:n}}{l(n)} \int_{\mathcal{X}} K(Y_{n+1}|x) \pi(dx|D_{j:n})$$

for  $j = 1, \dots, n(\mathbf{p})$ .

- (ii) Otherwise, the probability that customer  $n+1$  is new and is the  $j$ th largest among  $n(\mathbf{p}) + 1$  possible ranks is,

$$(13) \quad \frac{q_{j:n}}{l(n)} \int_{\mathcal{X}} K(Y_{n+1}|x) H(dx)$$

for  $j = 1, \dots, n(\mathbf{p}) + 1$ .

Similar to the gWCR SIS algorithms [see Ishwaran and James (2003, Lemma 2)], by appealing to the product rule of probability, repeating this procedure for customers  $\{1, \dots, n\}$ , produces a draw of  $\mathbf{m}$  from a density of  $\mathbf{m}$  depending on  $\mathbf{Y}$ , say  $q(\mathbf{m})$ , that satisfies the relationship

$$L(\mathbf{m})q(\mathbf{m}) = \pi_{\rho}(\mathbf{m}) \prod_{j=1}^{n(\mathbf{p})} \int_{\mathcal{X}} \prod_{i \in D_j} k(Y_i|x) H(dx)$$

where  $L(\mathbf{m}) = \prod_{i=1}^n l(i-1)$ . Hence for any functional,  $h(\mathbf{m})$  it follows that

$$(14) \quad \sum_{\mathbf{p}} \sum_{\mathbf{m} \in S_{n(\mathbf{p})}} h(\mathbf{m}) \pi_{\rho}(\mathbf{m}|\mathbf{Y}) = \frac{\sum_{\mathbf{p}} \sum_{\mathbf{m} \in S_{n(\mathbf{p})}} h(\mathbf{m}) L(\mathbf{m}) q(\mathbf{m})}{\sum_{\mathbf{p}} \sum_{\mathbf{m} \in S_{n(\mathbf{p})}} L(\mathbf{m}) q(\mathbf{m})}.$$

If the functional  $h(\mathbf{m})$  has a closed form, such as the predictive density  $\mathbb{E}[f(y|P)|\mathbf{m}, \mathbf{Y}] = f(y|\mathbf{m}, \mathbf{Y})$ , then one approximates (14) by using the rules in (12) and (13) to draw  $\mathbf{m}$ . Repeating this procedure say  $B$  times, results in iid realizations say  $(\mathbf{m}_{(b)})$  for  $b = 1, \dots, B$  and one can approximate (14) by

$$\frac{\sum_{b=1}^B h(\mathbf{m}_{(b)}) L(\mathbf{m}_{(b)})}{\sum_{b=1}^B L(\mathbf{m}_{(b)})}.$$

When the kernels  $K$  are set to 1, this procedure reduces to that described in James (2003, 2006) producing an exact draw from  $\pi_\rho(\mathbf{m})$ . For more intricate models one can incorporate a draw from the unique values  $X_1^*, \dots, X_{n(\mathbf{p})}^*$  which has the same distribution that arises for the Dirichlet process. One can also incorporate draws from the posterior distribution of  $P_{\rho, H}(dx)$  which is described in James (2006). Otherwise it is a simple matter to modify all the computational procedures discussed in Ishwaran and James (2003, section 4).

### 3.1 Normal Mixture example

One of the most studied and utilized Bayesian mixture models is the Normal mixture model, specified by the choice of

$$(15) \quad f_\sigma(y|P) = \int_{-\infty}^{\infty} \phi_\sigma(y-x)P(dx)$$

where

$$\phi_\sigma(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{z^2}{2\sigma^2}\right)$$

is a Normal density, which is a natural candidate for density estimation. In the case of the Dirichlet process, this model was introduced by Lo (1984) and popularized by the development of feasible computational algorithms in Escobar (1994) and Escobar and West (1995). Suppose that  $(Y_i)$  are iid with true density  $f_0$ , a recent result of Lijoi, Prünster and Walker (2005) shows that  $f_\sigma(\cdot|P)$  in (15) based on very general random probability measures, and a suitable prior distribution for  $\sigma$ , have posterior distributions that are strongly consistent in terms of estimating the unknown density  $f_0$  under rather mild conditions. In particular their result validates the use of rather arbitrary NTR species sampling models in this context with the classical choice of  $H$  set to be a Normal distribution with mean 0 and variance  $A$ . Here setting  $\sigma = \sqrt{\theta}$  one has

$$K(Y_i|X_i) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{1}{2\theta}(Y_i - X_i)^2\right).$$

Using these specifications we present the details of the proposed algorithm.

- (i) Customer  $n+1$  is seated to a new table and assigned rank  $j$  among  $n(\mathbf{p})+1$  possible ranks with probability

$$\frac{q_{j:n}}{\lambda_\theta(n+1)} \frac{1}{\sqrt{2\pi(\theta+A)}} \exp\left(-\frac{Y_{n+1}^2}{2(\theta+A)}\right)$$

- (ii) Customer  $n+1$  is seated to an existing table and is assigned rank  $j$  with probability

$$\frac{p_{j:n}}{\lambda_\theta(n+1)} \sqrt{\frac{\theta + Ad_j}{2\pi\theta[\theta + A(d_j+1)]}} \exp\left[-\frac{1}{2\theta} \left(Y_{n+1}^2 - \frac{A \sum_{i \in D_j} Y_i + Y_{n+1}}{\theta + A(d_j+1)} + \frac{A \sum_{i \in D_j} Y_i}{\theta + Ad_j}\right)\right]$$

- (iii) Additionally each  $X_j^*|\mathbf{Y}, \mathbf{m}, \theta$  is normally distributed with parameters

$$\frac{1}{\sigma_j} = \frac{d_j}{\theta} + \frac{1}{A} \text{ and } \mu_j = \frac{\sigma_j}{\theta} \sum_{i \in D_j} Y_i.$$

$\lambda_\theta(n+1)$  is the appropriate normalizing constant which is a special case of  $l(n)$ .

REMARK 3. For comparison, the setup and notation we use is similar to that used in Ishwaran and James (2003, 6.1) which is based on weighted Chinese restaurant sampling of partitions  $\mathbf{p}$ .



**4. Concluding Remarks.** We have given a brief account of how one can use Kjell Doksum's NTR models to create a new class of species sampling random probability measures which can be applied to complex mixture models. These models exhibit many features of the NTR models, in terms of clustering behavior, but as we have shown are simpler to use. Ideally one would like to describe parallel schemes for the more complex Spatial NTR models. However, this constitutes a considerably more involved study which we shall report elsewhere. More details can be found in James (2003, 2006) where explicit examples can be easily constructed.

The representation in (4) is important as it connects NTR processes to a large body of work on exponential functionals of Lévy processes which have applications in many fields including physics and finance. For a recent survey see Bertoin and Yor (2005). Some recent works which exploit this representation and are directly linked to NTR processes are Epifani, Lijoi and Prünster (2003) and James (2003, 2006). Additionally, outside of a Bayesian context, there is a notable body of recent work which has some overlaps with James (2003, 2006) and hence NTR processes by Gnedin and Pitman (2005a) and subsequent papers Gnedin and Pitman (2005b), Gnedin and Pitman and Yor (2005) and Gnedin, Pitman and Yor (2006). Although outside of a specific Bayesian context these papers contain results which are relevant to statistical analysis such as results related to the behavior of the number of ties  $n(\mathbf{p})$ . The fact that these models arise from different considerations and different points of emphasis attests to their rich nature. We are quite interested to see what future connections will be made.

## References

- ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152-1174.
- BERTOIN, J. AND YOR, M. (2005). Exponential functionals of Lévy processes. *Probab. Surv.* **2** 191-212.
- BLACKWELL, D. AND MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353-355.
- DOKSUM, K. A. (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2** 183-201.
- DOKSUM, K. A. AND JAMES, L. F. (2004). On spatial neutral to the right processes and their posterior distributions. In *Mathematical Reliability: An Expository Perspective*, Editors: Mazzuchi, Singpurwalla and Soyer. International Series in Operations Research and Management Science. Kluwer Academic Publishers.
- EPIFANI, I., LIJOI, A., AND PRÜNSTER, I. (2003). Exponential functionals and means of neutral to the right priors. *Biometrika* **90** 791-808.
- ESCOBAR, M.D. (1994). Estimating normal means with the Dirichlet process prior. *J. Amer. Stat. Assoc.* **89** 268-277.
- ESCOBAR, M.D. AND WEST, M. (1995 Bayesian density estimation and inference using mixtures.).
- EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3** 87-112.
- FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209-230.
- FERGUSON, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2** 615-629.
- FERGUSON, T. S. AND PHADIA, E. (1979). Bayesian nonparametric estimation based on censored data. *Ann. Statist.* **7** 163-186.
- FREEDMAN, D. A. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. *Ann. Math. Statist.* **34** 1386-1403.
- GNEDIN, A. V. (2004). Three sampling formulas. *Combin. Probab. Comput.* **13** 185-193.
- GNEDIN, A. V. AND PITMAN, J. (2005a). Regenerative composition structures. *Ann. Probab.* **33** 445-479.
- GNEDIN, A. V. AND PITMAN, J. (2005b). Self-similar and Markov composition structures. In *Representation Theory, Dynamical Systems, Combinatorial and Algorithmic Methods*. Part 13, A. A. Lodkin editor. Zapiski Nauchnyh Seminarov POMI, Vol. 326, PDMI, 59-84.

- GNEDIN, A. V. AND PITMAN, J. AND YOR, M. (2005). Asymptotic laws for regenerative compositions: gamma subordinators and the like. *Probab. Th. and Rel. Fields. Published online November 2005*
- GNEDIN, A. V. AND PITMAN, J. AND YOR, M. (2006). Asymptotic laws for compositions derived from transformed subordinators. *Ann. Probab.* **34**
- HJORT, N. L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Ann. Statist.* **18** 1259-1294.
- ISHWARAN, H. AND JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96** 161-173.
- ISHWARAN, H. AND JAMES, L. F. (2003). Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica* **13** 1211-1235.
- JAMES, L. F. (2003). Poisson calculus for spatial neutral to the right processes(Big version). arXiv:math.PR/0305053. Available at <http://arxiv.org/abs/math.PR/0305053>.
- JAMES, L. F. (2006). Poisson calculus for spatial neutral to the right processes. *Ann. Statist.* **34**
- KIM, Y. (1999). Nonparametric Bayesian estimators for counting processes. *Ann. Statist.* **27** 562-588.
- LIJOI, A., PRÜNSTER, I. AND WALKER, S.G. (2005). On consistency of nonparametric normal mixtures for Bayesian density estimation. *J. Amer. Stat. Assoc.* **100** 1292-1296.
- LO, A. Y. (1993). A Bayesian bootstrap for censored data. *Ann. Statist.* **21** 100-123.
- LO, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density Estimates. *Ann. Statist.* **12** 351-357.
- LO, A.Y., BRUNNER, L.J. AND CHAN, A.T. (1996). Weighted Chinese restaurant processes and Bayesian mixture model. Research Report Hong Kong University of Science and Technology.
- MÜLLER, P, AND QUINTANA, F. A. (2004). Nonparametric Bayesian data analysis. *Statist. Sci.* **19** 95-110.
- PITMAN, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In Statistics, Probability and Game Theory T.S. Ferguson, L.S. Shapley and J.B. Macqueen editors, IMS Lecture Notes-Monograph series, Vol 30, pages 245-267.
- PITMAN, J. AND YOR, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25** 855-900.
- WALKER, S. AND MULIERE, P. (1997). Beta-Stacy processes and a generalization of the Pólya-urn scheme. *Ann. Statist.* **25** 1762-1780.

LANCELOT F. JAMES  
 THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY  
 DEPARTMENT OF INFORMATION SYSTEMS AND MANAGEMENT  
 CLEAR WATER BAY, KOWLOON  
 HONG KONG  
[lancelot@ust.hk](mailto:lancelot@ust.hk)